Geo-Tagging for Imprecise Regions of Different Sizes

Robert C Pasley
University of Sheffield
Western Bank
Sheffield, UK
+44 (0) 114 2222666
r.pasley@sheffield.ac.uk

Paul Clough
University of Sheffield
Western Bank
Sheffield, UK
+44 (0) 114 2222664
p.d.clough@sheffield.ac.uk

Mark Sanderson
University of Sheffield
Western Bank
Sheffield, UK
+44 (0) 114 2222648
m.sanderson@sheffield.ac.uk

ABSTRACT
Extracting geographical information from various web sources is likely to be important for a variety of applications. One such use for this information is to enable the study of vernacular regions: informal places referred to on a day-to-day basis, but with no official entry in geographical resources, such as gazetteers. Past work in automatically extracting geographical information from the web to support the creation of vernacular regions has tended to focus on larger regions (e.g. "The British Midlands" and "The South of France"). In this paper we report the results of preliminary work to investigate the success of using a simple geo-tagging approach and resources of varying granularity from the Ordnance Survey to extract geographical information from web pages. We find that the data gathered for smaller regions (compared with larger ones) is more "fine-grained" which has an effect on the type of resource most useful for geo-tagging and its success.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval]**

General Terms
Algorithms, experimentation.

Keywords
Vernacular regions, web mining, geo-tagging.

1. INTRODUCTION
As noted in [3], [24] and [13] there is a semantic gap between the requirements of Geographical Information Systems (GIS) users and the functionality supported by these systems. GIS tend to allow access to spatial information in a spatial way, using primitives such as points, lines and polygons. However, there is relatively little support for the use of place names. Waters and Evans [35] point out that although people do not tend to use a scientific geographical vocabulary, they do tend to use many geographical terms on a "day to day basis".

GIS'06, November 10–11, 2006 Arlington, Virginia, USA.

Waters and Evans [35] cite two examples, "downtown" and the "grim area around the docks". The type of regions described by these terms are referred to as vernacular geography and these kinds of references are of a type not commonly contained within gazetteers that are used with GIS.

Gazetteers containing vernacular references would be useful within many computer applications. For example, according to a study of query logs by Kohler and Sanderson [29], users of web search engines have a tendency to issue queries with a geographical dimension. Online information services such as Google Maps (local search) and Multimap (mapping service) would benefit from a gazetteer of vernacular geography enabling users to incorporate vernacular geography terms into their search (e.g. "hotels in the British Midlands") and national mapping agencies such as the Ordnance Survey in Britain could enhance existing resources with non-administrative information. Gazetteers such as these may be useful within Geographical Information Retrieval (GIR) [14], marketing, the supply of culturally dependant services, emergency services, online mapping services and possibly even as input to boundary re-assignation decisions.

The aim, then, is to somehow gather information which can be used to detect and model vernacular geography regions (referred to here as *imprecise regions*). Previous work has attempted to define vernacular areas such as "downtown Santa Barbara" [27], "'high crime areas' in the city of Leeds" [35] and "Sheffield City Centre" [22] through manual data collection methods. More automated techniques have been initially tested based on gathering and mining data from the web [15] [2]. However, unlike the manual approaches, these have only been used for larger regions (e.g. "The British Midlands" and "Mid Wales").

This paper describes our initial work in using sources gathered from the web to inform the generation of imprecise regions for varying size (or granularity). In order to achieve this, suitable web pages must be identified and checked for geographical references. Sources such as Wikipedia, blogs and discussion forums may have a wealth of information including geographical definitions, both administrative and vernacular. The information found there may be useful in building vernacular gazetteers, however to-date no work has investigated web mining for imprecise regions of varying size (e.g. "The British Midlands" versus "Hunter's Bar in Sheffield"). This paper investigates the impact of region size on the success of extracting geo-references from web pages and shows that the source of errors differs when dealing with smaller regions.

2. BACKGROUND

2.1 Web Mining

Text mining addresses the extraction of data from documents using shallow parsing techniques. Web mining is a branch of text mining concerned with the web; it tends to be more challenging as the semantics of web-pages are not as predictable as, for example, newswire text. According to [11] web pages are more complex and dynamic than traditional text sources as well as serving a broader spectrum of communities. Web Mining can be approached in several ways. One way to look at the web is as an essentially linguistic resource [16]. This approach pays less attention to the HTML structure and assumes that natural language processing (NLP) techniques can be used on the text as it is displayed on the screen. Alternatively, with web pages that are essentially structured the same as each other, but with different data, these data can be extracted and formed into rows in a database [10] e.g. lists of hotels. In the experiment in this paper the HTML structure is preserved in order to allow the pages to be displayed in a browser; however for the purposes of tagging the structure is ignored thus treating the web as an essentially linguistic medium [16].

2.2 Identifying & Using Geo-references

Extracting geospatial information from web pages involves two main tasks (collectively referred to as geo-tagging). Firstly identifying geo-references (e.g. place names, addresses, address fragments, postcodes and telephone numbers) commonly referred to as geo-parsing (or geo/non-geo disambiguation) [5, 17] and assigning them spatial coordinates (a point, line or polygon) commonly referred to as geo-coding (or geo/geo disambiguation) [1, 8]. Geo-coding might also include identifying or defining the geographical scope of a document. There are many approaches for automatically identifying and grounding geo-references which have been used mainly on newspaper texts and web pages. An overall aim of our work is to evaluate and improve techniques for dealing with the wide variety of information which can be found on the web (e.g. blogs, discussion forums and wikis) and likely to contain useful informal geographical information.

It is worth noting that existing techniques are typically used to identify geo-references at a particular granularity (e.g. at a level of city or town), on fairly specific types of data (e.g. web pages and newspaper texts) and using readily available geographical resources. In this work we aim to address a much lower level of granularity (e.g. street-level), with varying types of web data (blogs, home pages, directories, Wikis and discussion forums) and using a range of different geographical resources as provided by the Ordnance Survey (see Section 3.3).

2.3 Evaluation

As Leidner [19] points out, evaluation on toponym resolution (TR) is inconsistent. He presents evaluation at a particular level of granularity (that of roughly town level upwards) and only includes populated places (i.e. not topographic places such as airport, bridges etc.). It is also assumed that geo/non-geo disambiguation has already been done; the systems would be evaluated on the basis of discovering which place the toponym refers to, not whether it is a toponym or some other type of word. A gold standard test set is created, which consists of passages of news text where all toponyms have been annotated. Once this gold standard was established it was possible to reconstruct some of the most promising TR systems and test them against each other, something that was hard to do previously. Clough and Sanderson [7] also highlight the need for creating a standardised resource for the evaluation of geo-tagging, particularly on types of text such web pages.

3. EXTRACTING GEO-REFERENCES

3.1 Geo-parsing

To extract geographical information from web pages, allow the manual annotation of example data and perform the evaluation of automated extraction techniques, the General Architecture for Text Engineering (GATE) system has been used. GATE provides a Collection of REusable Objects for Language Engineering (CREOLE), a set of general resources integrated into GATE [10]. The CREOLE consists of resources such as ANNIE (A Nearly New Information Extraction system), a default Information Extraction (IE) system, which includes a tokeniser, gazetteer manager, sentence splitter, part-of-speech tagger, semantic tagger and a co-reference module. Many functions essential to NLP are provided in an easy to use interface, allowing quick access to functions such as gazetteer lookup, annotation interface, grammar rules which work with annotations, and an annotation comparison sub-system. GATE can be run from its own interface or incorporated using standard libraries into Java programs. In this experiment a Java program was used to create the machine annotations so that access to a MySQL database containing OS data (see Section 3.3) was possible. This approach would allow various databases to be created in the future, perhaps allowing other ontologically based gazetteers to be used. The manual annotation was done using the GATE interface.

3.2 Geo-coding

Similar to [33] the geo-coder in this experiment works as follows: following the identification of possible geo-references (Section 3.1), each geo-reference is compared with those in the MySQL database containing all of the OS data as outlined below (Section 3.3). All entries that match geo-references for the web page are extracted from the database. The northing (x) and easting (y) co-ordinates are averaged to derive the centroid of all the *possible* groundings of all the geo-references found in the web page. This centroid can be thought of as an average point for the full set of potential geo-references. For each geo-reference the possible co-ordinates closest to the centroid point are selected as the most plausible location of that geo-reference for this document. For example, a document that mentioned Sheffield and Barnsley would use Sheffield, South Yorkshire whereas a document mentioning Sheffield and Mousehole would use the Sheffield in Cornwall.

3.3 Geographical Resources

The Ordnance Survey has kindly provided various data sources for our research. These resources provide various representations of places in geographical terms; however the detailed extents of places are not available within these resources. The resources are:

- **[os50k] 50k Gazetteer**: this lists places that appear on the 1:50000 OS maps (e.g. populated places and certain landmarks). The places are geo-coded with co-ordinate *points* for the kilometre squares appropriate to the places.

- **[osl] OS Locator**: this contains street names. The streets are sectioned up as this makes the data clearer and spatial data consists of a *minimum bounding box* (MBB) for the section and a representative point within the MBB. There is no postcode information since the sections can span multiple postcodes.
- **[oscp] Code-Point** with Polygons: this data contains all postcodes and provides a set of polygons for each postcode.
- **[osmm] OS Master Map**: address layer2, part of osmm, contains all addresses in a Postal Address Format (PAF); which is an agreed address standard. Each address also has a co-ordinate *point*. The data representing the South West and Centre of Sheffield was available to us for this experiment.

When attempting to determine the shape of vernacular regions from Web data, it may be for that some regions there are many web data sources; for others there is less coverage. Some types of geo-reference are more ambiguous than others. The address level could be more ambiguous than toponyms; however if complete addresses can be geo-coded it may be found that they have a lower level of ambiguity. It should be possible to geo-code using phone area codes since they have a geographical scope. However, no resources were available to conduct such coding.

The OS master map address level 2 contains a large proportion of known UK addresses; this data is geo-coded with OS co-ordinates. Addresses would have to be "fuzzy matched", since we could not expect the addresses on web sites to be complete and labelled to the standardised PAF format. This data could be expected to have a high degree of ambiguity at the partial match level. The next level of granularity has all UK postcodes, with polygon data. Working with this data set might impose the need to geo-code within the spatial area of the postcode, which might be vulnerable to inaccuracies. Another available level holds the street name data; this has address information at locality and settlement level, but does not have postcode data because streets can span many postcodes. There is also a bounding box for the extent of the street and a representative point. In order to keep the bounding boxes minimal the street is often cut up into sections in the data.

4. EXPERIMENTAL SETUP

4.1 Experiment

As an initial investigation into defining vernacular regions using information from the Web, an experiment was carried out in order to assess the differences made by granularity of geographic scope of web sites in relation to the various Ordinance Survey resources. Firstly, pages known to be about particular vernacular regions were collected. Three regions were chosen with a range of physical sizes.

- **The Midlands:** this is a region in Central England thought to be about 130km x 150km (derived from [22]).
- **Sheffield City Centre**: is just the central part of Sheffield, South Yorkshire, UK thought to be about 1km x 1½km (derived from [22]).
- **Hunters Bar:** (an area in Sheffield) judged to be about ½km x ½km by the authors of the paper.

Web pages for these regions were collected manually by using handcrafted queries such as "Midland Cities –east –west" using Google UK. The searcher verified that the web sites harvested in this manner were relevant to the target region. Once the set of relevant pages had been collected all geographical references were extracted (section 3.1) and grounded (section 3.2). The web pages then represented a set of associations between the target vernacular regions and the geographical references.

4.2 Evaluation Data

Using the benchmark data we evaluated the geo-parsing methods using the default version of ANNIE from GATE and gazetteer lookup. We experimented with using each geographical resource and for each system setting we used the GATE AnnotationDIFF tool to compare the benchmark annotations (key-set) with those generated by the system (response-set). AnnotationDiff was unable to automatically provide statistics for the differences because human judgment was required to check that the annotations differed or not. "Winter Garden", a place in Sheffield, and "Garden" should not be a partial match, but "Sheffield Town Centre" and "Sheffield" could be viewed as a partial match.

In these experiments we have been provided with large and comprehensive data sources, and expect to process a large number of documents, it is important to maximize efficiency in this project. Therefore these reasons the current system uses a simple gazetteer lookup approach. Although simple, this approach is robust which is important as web data is often "noisy" (i.e. ungrammatical). Clough used a similar approach for web pages in [6]. All manual annotations are created with respect to the database of MySQL, viewed through a browser-based client, designed specifically for the purpose.

Geo-references were tagged with an attribute to signify which resource item matched the text; those referred to as "Not in resources" in the results are the ones where no resource item could be assigned to the text.

It was observed during the experiment that line features such as roads and rivers often appear multiple times in the OSL data. This is because multiple points are needed in order to represent their existence along their length. Conversely hamlets, towns, villages and cities are all represented by a point somewhere close to their centre and give no clue to the actual extent of the place represented. The way many features are represented in the data often caused problems for the manual annotation process, it would be useful to be able to have available an object that represented more specifically the extent of a feature. It should also be noted that it is hard to annotate addresses. One has to decide whether to annotate each part of the address, or attempt to annotate the whole address at once to a specific point.

Table 1. Geo-tagging effectiveness for regions of different sizes

Region		Correct		Error	
	Geo-refs		False positives	Incorrect	Incorrect gazetteer
The Midlands	2523	21.2%	50%	28.7%	26.1%
Sheffield City Centre	3503	17.7%	30.9%	51.4%	40%
Hunters Bar	4037	17.5%	25.0%	57.5%	47.7%

5. RESULTS

Table 1 shows the percentage of correct annotations, those where text was marked as geographical when the manual annotator did not mark it (*false positives*), and those either not marked or marked wrongly (*incorrect*). These have been combined because often the extent of the annotations differ, thus making it hard to compare the two sets of annotations. The last figure (*incorrect gazetteer*) is split down into references that were purely geographic and might be expected to exist in a standard gazetteer, and those, such as phone numbers and building names, that would require knowledge from elsewhere.

Although the *correct* column is similar for the three regions, it can be seen that the source of errors differ. The low *correct* result is not unexpected since the geo-tagging approaches used in this experiment are the simplest which could be expected to give useful results. The *error* columns suggest that improvements on this basic tagger would be made differently depending on the type of web-pages and the geographical scope that those web-pages had. Where the scope is large improvements would be derived from reducing the false positive, and on smaller area improvements would come from identifying and geo-coding references better. It should also be noted that where indirect geographical references are used, such as telephone area codes and names of landmark buildings, the number of these incorrectly recognised increases as the size of the scope reduces. This may be due to a larger number of such references in the web-pages.

Many of the false positives come from surprising entries in the gazetteer. The gazetteers have entries that include words such as "Banks", "Garden", "1" and "Society", all parts of very common place names and text common in most types of document. The other resources are more geographically based, containing postcodes and addresses.

Table 2. Matches between manually-annotated pages and OS resources

%	os50k	osl	oscp	osmm	Not in resources	*Total*
The Midlands	47.3	17.8	1.9	4.3	28.7	*25*
Sheffield City Centre	38.2	14.2	2.5	3.9	41.2	*34.8*
Hunters Bar	47.3	19.5	2.5	5.4	25.3	*40.2*
Total	*44.1*	*17.1*	*2.3*	*4.6*	*31.7*	*100%*

By observation of the 10,063 geo-references in the manually annotated pages it was noted that (Table 2): Sheffield City Centre contains many references that could not be geo-tagged because they were names of landmarks and Hunters Bar contained many more references at a lower level of granularity than Midlands or Sheffield City Centre (shown in Table 2) by the increase in the numbers of geo-references from oscp and osl - osmm coverage was limited for The Midlands.

6. DISCUSSION

It was found that the source of errors changed as the granularity reduced. The resources needed to geo-tag the smaller regions are finer-grained, such as addresses. False positives reduce as the scope reduces, and so are less of a problem. The resources with the largest places are least suitable for the web-sites with the smallest region of scope. This seems to suggest that address matching will increase accuracy with which address text and the rest of the document, especially at the higher resolution of granularity, is defined. Extracting postcodes, addresses and telephone numbers relies more on pattern matching than on disambiguation. This is because these tend to be less ambiguous than toponyms [33].

The geo-tagger used is of a relatively simple design. Due to the techniques used geo-references that are highly ambiguous overwhelm ones that are not in the calculation of the centroid. This is counter-intuitive and should be eliminated by a weighting scheme. It should also be noted that multiple mentions of the same place name skew the centroid towards that place name. This may be appropriate since where points appear multiple times it might be that they are deemed to be more important to the scope (more central) than those that appear few times [36]. The proximity of place names in document may infer that assumed referents closest to unambiguous places are more plausible. [20, 28], this would imply that the unambiguous places should be weighted more strongly.

Also if a geo-reference is found in the web-page but the appropriate referent (place) is not in the database, an inappropriate one would always be used; e.g. if Perth, Australia should not be geo-coded as Perth, Scotland. It may be possible to add more world knowledge to the system, and to develop a technique similar to [2] to allow places outside of the UK to be chosen in disambiguation, but to then be rejected as places in the definition of the extent of the imprecise region (since it is unlikely to be related to the region in question).

Since the Ordnance Survey data sets were not designed as a knowledge source for web-mining they need to be augmented from other sources such as Wikipedia, blogs and forum sites. Investigations into the feasibility of creating an ontology of vernacular geography should be made [26]. Another possibility is to build a co-location dictionary to allow for better disambiguation using the textual context, another would be the addition of extents for places in the gazetteer.

It should be acknowledged that not all geo-tagging could be done using one method since the depth of any world knowledge could vary amongst the gazetteer entries. Therefore the most preferred methods should be used first and if these do not work gradually less reliable methods are used. [19, 18, 21]. The fall back method is to choose some criteria such as population size, level in hierarchy and default to one of the limits [2]. A default can be chosen on the basis of distance in an ontology tree, parentage, or peer similarity. Note that each geo-reference in a document is assumed to be the same throughout the document or set of documents [31]. Ambiguity in large gazetteers can be avoided by filtering entries through Wikipedia to find a default place [4].

It is envisaged that a guided crawler will be created. Using a technique such as pseudo-visual prioritisation (such as the VIPS system [37, 5]) links will be chosen which show the greatest promise of being similar or relevant pages.

Contextual pattern matching (or trigger-phrase), e.g. searching for the regular expression "* city", should yield * as a geo-reference [2]. These worked well, however they are unlikely to have found all possible patterns. Statistical methods could be used to train a system that can learn co-occurrence patterns using a training

corpus to spot context that implies a place has been mentioned (rather than a person or organisation) [21] [12].

Statistical methods rely on having sufficient training cases available. It may be possible to gather these by bootstrapping for gazetteer entry collection. E.g. if we know some cities we could use those names to collect the sort of sentences in which city names occur, these could be used in a similar way to the trigger phrases above to collect entries for the gazetteer. [25, 32]. A useful possible side effect of this technique would be that the context could be used independently of the actual place name, this would yield a list of previously unknown place names.

Domain knowledge in the form of structured and feature rich gazetteers can be used to disambiguate references, e.g. "Hastings city" would not make sense (in England) because Hastings is not a city [23]. Reasoning could be carried out using relationships such as "X is north of Y" extracted from knowledge. [30, 9, 20, 28]. This knowledge could be collected by web-mining e.g. in [26], the authors investigated using Wikipedia to derive specialist thesauri.

A survey could be undertaken in an attempt to understand people's perceptions of vernacular regions. This would allow empirical evaluation of the effectiveness of any techniques that are used. It may be possible to use a web-site such as BBC Voices to publicize this survey which could be undertaken by a large population sample via the web. Sites such as BBC Voices, YouTube and Flickr allow user generated content, and there may be important clues in the forums and comments of such web pages. Blogs have become popular and are often geo-tagged. It should be possible to use these tags directly, and since they are co-ordinates they are already disambiguated and geo-coded. Some of the language used on these sites is colloquial, but the content may be more valid for this work than the more official, better written sites. Since NLP is typically less effective on colloquial text than on well-written text, robust techniques need to be developed for these sites.

The number of regions used in this experiment is limited. For example there are no regions between "The Midlands" and "Sheffield City Centre" which are markedly different in size. A later experiment using more regions to give a better sample size would allow the effect of granularity to be seen more clearly. The experience gained in this experiment should be used to improve the annotation scheme and increase the level of automation in the geo-tagging and evaluation processes, thus allowing easier experimentation.

It is envisaged that it will be possible to use NLP techniques to extract geo-references such as postcodes, addresses (full and partial), well-known landmarks, toponyms and telephone area codes from text passages in web-pages, as well as any geo-tags that might exist. It is also envisaged that structured and semi-structured web pages will be able to provide full addresses. These semi-structured pages would contain lists of relevant items such as "hotels in the Midlands".

CONCLUSIONS

We find that the data gathered for smaller regions (compared with larger ones) is more "fine-grained" which has an effect on the type of resource most useful for geo-tagging and its success.

ACKNOWLEDGMENTS

Work partially supported by the EPSRC and Ordnance Survey (CASE/CAN/06/67). The authors are solely responsible for the content of this paper. It does not represent the opinion of supporting funding agencies, and the supporters are not responsible for any use that might be made of data appearing therein.

REFERENCES

[1] Amitay, E.; Har'El, N.; Sivan, R. and Soffer, A. "Web-a-where: Geotagging Web Content". In *Proceedings of the SIGIR 2004 conference on Research and Development in Information Retrieval* (Sheffield, UK, Sheffield, July 25th - 29th 2004). ACM Press, New York, NY. 2000. 273-280.

[2] Arampatzis, M., Reinbacher, I., Jones, C., Vaid, S., Clough, P., Joho, H. and Sanderson, M. Web-based delineation of imprecise regions. *Journal of Computers, Environments and Urban Systems (CEUS)*, 30(4). 2004. 436-459.

[3] Burrough, P.A. & Frank, A.U. Concepts and Paradigms in Spatial Information: Are Current Geographical Information Systems Truly Generic?. *International Journal of Geographical Information Systems.* 9(2). 1995. 101-116.

[4] Buscaldi, D., Rosso, P. & García, P.P. Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval IR In *3rd Workshop on Geographic Information Retrieval (GIR 2006)*. Seattle, WA, USA. 2006.

[5] Cai, D.; Yu, S.; Wen, J. & Ma, W. Extracting Content Structure for Web Pages Based on Visual Representation. In *Fifth Asia Pacific Web Conference (APWeb-03).* 2003. 406-417.

[6] Clough, P. Extracting metadata for spatially-aware information retrieval on the internet, *In GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*. ACM Press, New York, NY, USA. 2005. 25-30.

[7] Clough, P. & Sanderson, M. A proposal for comparative evaluation of automatic annotation for geo-referenced documents, In *Workshop on Geographic Information Retrieval* held at the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, page unnumbered. Association for Computing Machinery, Sheffield, England, UK. 2004.

[8] Cunningham, H. GATE, a General Architecture for Text Engineering. *Computers and the Humanities* 36. 2002. 223-254.

[9] Egenhofer, M. Toward the semantic geospatial web. In proceedings of ACM-GIS 2002. 2002. 86-95.

[10] Embley, D.; Jiang, S. & Ng, Y. Record-boundary discovery in Web documents. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, Philadelphia, Pennsylvania, United States. 1999. 467 – 478.

[11] Han, J., and Chang, K.C. *Data Mining for Web Intelligence*. IEEE Computer, IEEE Computer Society, Washington, D.C. 35(11). 2002. 64-70.

[12] Hartrumpf, S. & Leveling, J. On Metonymy Recognition for Geographic IR In *3rd Workshop on Geographic Information Retrieval (GIR 2006).* Seattle, WA, USA. 2006.

[13] Heinzle, F., Clough, P., Elias, B., Sester, M. Metadata Annotation Methods (d29 6301). SPIRIT Project technical report. 2005.

[14] Hill, L. L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries* (September 18 - 20, 2000). J. L. Borbinha and T. Baker, Eds. Lecture Notes in Computer Science, vol. 1923. Springer-Verlag, London. 2000. 280-290.

[15] Iria, J. & Ciravegna, F. Relation Extraction for Mining the Semantic Web. In: *Dagstuhl Seminar: Machine Learning for the Semantic Web*. (13-18 Feb 2005) Germany.

[16] Kilgarriff, A. & Grefenstette, G. Introduction to the special issue on the web as corpus, *Comput. Linguist.* 29(3). 2003. 333-334

[17] Larson, R. Geographic information retrieval and spatial browsing. *GIS and Libraries: 32nd Annual Clinic on Library Applications of Data Processing Conference* (Urban-Champaign: University of Illinois), 1996. 81–124.

[18] Leidner, J. L. Toponym Resolution in Text: `Which Sheffield is it?' In *Proceedings of the SIGIR 2004 conference on Research and Development in Information Retrieval (Sheffield, UK,* Sheffield, July 25th -29th 2004). ACM Press, New York, NY. 2004. 602.

[19] Leidner, J. L. An Evaluation Dataset for the Toponym Resolution Task. Computers, Environment and Urban Systems Special Issue on Geographic Information Retrieval. 30(4). Elsevier Science. 2006. 400-417.

[20] Li, S. & Ying, M. Extensionality of the RCC8 composition table. Fundamenta Informaticae, 55(3-4). 2003. 363 – 385.

[21] Li, Y.; Moffat, A.; Stokes, N. & Cavedon, L. Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In *3rd Workshop on Geographic Information Retrieval (GIR 2006).* Seattle, WA,USA, 2006. 17-22.

[22] Mansbridge, L. *Perceptions of Imprecise Regions in Relation to Geographical Information Retrieval*. MSc Thesis, University of Sheffield. 2005.

[23] Martins, B.; Silva, M.J.; Freitas, S. & Afonso, A.P. Handling Locations in Search Engine Queries. in *Proceedings of GIR-2006, the 3rd Workshop on Geographical Information Retrieval (held at SIGIR 2006).* 2006

[24] Mennis, J.L. Derivation and implementation of a semantic GIS data model informed by principles of cognition. *Computers, Environment and Urban Systems* 27(5). 2005. 455-479.

[25] Mikheev, A. Moens, M. & Grover, C. Named Entity Recognition without Gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL).* Bergen, Norway. 1999. 1-8.

[26] Milne, D. Medelyan, O. and Witten, I.H. Mining Domain-Specific Thesauri from Wikipedia: A case study. In *ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06).* 2006. 442-448.

[27] Montello, D. R., Goodchild, M. F., Gottesegen J., and Fohl, P. *Where's downtown? Behavioral methods for determining referents of vague spatial queries.* Spatial Cognition and Computation, 3(2&3). 2003. 185—204.

[28] Rauch, E.; Bukatin, M. and Baker, K. A confidence-based framework for disambiguating geographic terms, In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references.* Association for Computational Linguistics, Morristown, NJ, USA. 2003. 50-54.

[29] Sanderson, M. & Kohler, J. *Analyzing Geographic Queries.* MSc Thesis, University of Sheffield. 2004.

[30] Schockaert, S.; De Cock, M. and Kerre, E.E. Towards Fuzzy Spatial Reasoning in Geographic IR Systems. In *Proceedings of the 3rd IEEE International Conference on Intelligent Systems (IEEE IS 2006).* 2006. 221-226

[31] Silva, M.J.; Martins, B.; Chaves, M.; Afonso, A.P. and Cardoso, N. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban System.* 30. 2006. 378-399.

[32] Smith, D.A. & Mann, G.S. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, Association for Computational Linguistics, Morristown, NJ, USA. 2003. 45-49.

[33] Smith, D.A. & Crane, G. Disambiguating Geographic Names in a Historical Digital Library. Lecture Notes in Computer Science 2163. 2001. 127

[34] Wang, C.; Xie, X.; Wang, L.; Lu, Y. and Ma, W. Detecting Geographic Locations from Web Resources. In *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*. ACM Press, New York, NY, USA. 2005. 17-24.

[35] Waters, T. and Evans, A.J. Tools for the web-based GIS mapping of "fuzzy" vernacular geography Paper presented at GISRUK 2003, City University, London, (April 2003). 9-11.

[36] Woodruff, A.G. & Plaunt, C. GIPSY: Georeferenced Information Processing System (S2K-94-41), 2003. 24.

[37] Yu, S.; Cai, D.; Wen, J. & Ma, W. Improving pseudo-relevance feedback in web information retrieval using web page segmentation, In *WWW '03: Proceedings of the 12th international conference on World Wide Web*. ACM Press, New York, NY, USA. 2003. 11-18.